論文
9-5-6

자료 통계 분석을 위한 MS 엑셀의 유용한 기능들에 관한 사례연구 (지하철 이용객 자료 분석)

A case study of MS Excel's powerful functions for statistical data analysis. (Focused on an Analysis of Variance menu)

김 숙 영(Sook Young Kim)[1)]

요약

엑셀이 자료 통계 분석에서 매우 편리하고 유용한 도구가 될 수 있음을 보여주기 위하여 지하철 이용객 자료로 가설을 검정하는 사례 연구를 시행하였다.

양적 자료는 엑셀의 기술통계량 메뉴에 의하여, 질적 자료는 히스토그램 메뉴에 의하여 기술되었고, 변수들간의 관계성 검정은 회귀 분석 메뉴에 의하여, 차이 검정은 T검정 메뉴에 의하여, 요인 검정은 분산배치법 메뉴에 의하여 전문적인 결과를 얻을 수 있었다.

엑셀만을 이용하여 자료 입력, 관리 및 통계 분석 결과를 편리하게 수행할 수 있는 사례가 되었다.

Abstract

A case study to show MS Excel's convenient and powerful functions was conducted to test hypotheses with subway data.

Quantitative variables were described using descriptive menu, and qualitative variables were described using histogram menu of a MS Excel software.

Relationships were tested using regression menu, differences were tested using t-test menu, and factors were tested using variance-layout menu of a Excel software.

Data input, management, and statistical analysis were done successfully with only a MS Excel software..

Key words: Analysis of Variance, Regress, Excel functions, Statistics

논문 접수 : 2008. 12. 10.
심사 완료 : 2008. 12. 19.

1) 정회원 : 안산공과대학 컴퓨터 정보과

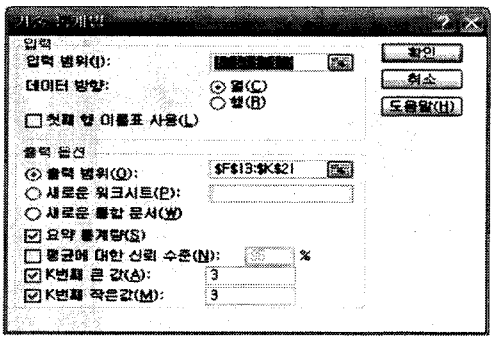
1. 서론

90년대 이후 정보화 사회에서는 연구분야를 막론하고 실제 정보로 연구가설을 논리적으로 입증하는 방법론을 적용하고 있다.
따라서 자료를 수집하고 통계적인 분석절차를 통하여 가설의 채택 여부를 결정하는 확률적 기법이 필수 도구이다.
특히 통계학의 중요성이 강조되는 몇 개 분야에서는 경제 통계학(Econometrics), 심리 통계(Psychometrics) 및 생물 통계학(Biostatistics) 등의 독자적인 학문을 이루고 있다.
수치적인 통계량 결과를 얻기 위하여는 확률분포 계산이 필요하므로 통계 전산 프로그램 사용이 필수적이며 대부분 통계 패키지를 사용하는 실정이다.
현재 많이 알려진 통계 패키지에는 사회과학 분야 연구에 유용하도록 개발된 SPSS(Statistical package for social science), 의학 분야 연구에 유용하도록 개발된 BMDP(Biomedical package) 및 특정 분야에 관계없이 강력한 기능들을 가진 SAS(Statistical Analysis package) 등이다.
대부분 수집된 정보를 엑셀등의 스프레드시트 소프트웨어로 자료 파일을 생성하고 통계 패키지에서 자료를 읽어 분석하는 절차를 채택하고 있다. 또한 보고서 작성을 위하여 또다시 엑셀등의 그래픽 기능을 이용해야 하는 복잡한 절차를 채택하고 있다.
통계 패키지 프로그램은 프로시저를 불러 사용하는 것 이므로, 프로시저를 호출하는 문법 이해와 또한 전문적인 결과들 까지도 출력되므로 필요한 결과 부분들만을 선택하는 설명도 필요하다. 자료 성격에 따라 변수들을 통합하는 과정 및 두 개 이상의 자료 파일을 사용해야 하는 경우도 있다.
이러한 데이터 관리 작업 및 연구 분야에 적합한 전문적인 차트 작성 작업들은 스프레드시트 프로그램인 엑셀에서 쉽게 수행될 수 있으므로, 통계 분석 결과도 엑셀에서 얻을 수 있으면 전문 통계 패키지 없이도 전문적인 통계 분석을 쉽고 편리하게 수행할 수 있다.
엑셀에 존재하는 데이터 분석 메뉴에는 모집단을 대표하는 표본을 추출하는 표본추출로부터 전문적인 결과를 제공하는 프로시저들을 포함하고 있다.
이와같이 엑셀을 이용하면 편리하고 효율적으로 통계 분석 결과 및 보고서를 작성할 수 있음에도, 통계 분석은 전문 통계 패키지를 사용해야 한다는 기존 개념에 의하여 엑셀에 존재하는 자료분석 메뉴 활용도가 매우 저조한 실정이다.
따라서 본 연구에서는 지하철 이용객 자료를 엑셀의 데이터 분석 메뉴들만을 사용하여 분석하는 사례연구를 통하여 엑셀의 통계 분석 도구의 유용함을 설명한다.

2. 조사 방법

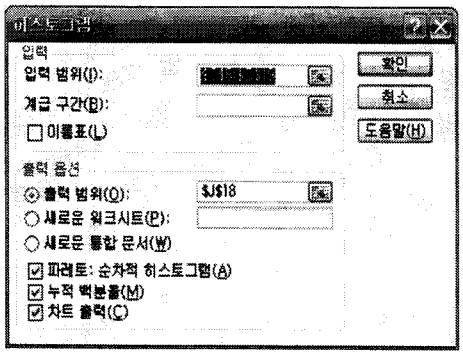
2-1. 이론적 배경

모든 통계 분석 첫 단계로 자료들을 기술하는 기술통계 메뉴이다.



<그림 1> 기술통계량 메뉴 대화상자

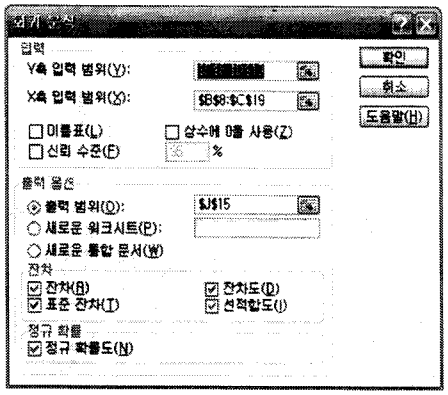
기술통계량 결과에서는 양적 자료에 대한 평균, 오차, 중앙값,등의 기본 통계량 수치 및 분포에 관한 정보를 제공하는 외도, 첨도등이 출력된다. 특히 신뢰수준 결과는 다른 통계 프로그램에서는 쉽게 얻을 수 없는 정해놓은 유의수준에서 평균값의 신뢰수준이 출력된다.[1]



<그림 2> 히스토그램 메뉴 대화상자

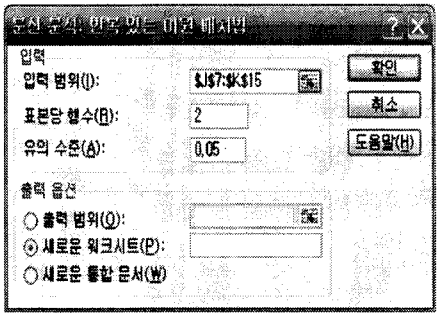
질적 자료 기술통계에 필수적인 히스토그램 메뉴에서는 계급에 속하는 빈도수 정보를 테이블과 그래프로 출력하고 있다. 분포가 균등한 자료에서는 계급값을 정하지 않아도 자동적으로 계급이 정하여져 결과를 제공하므로 매우 편리하다 (그림 2).

변수의 관계를 분석하는 엑셀 메뉴에는 상관분석과 회귀분석이 있다. 상관분석은 상관 계수의 값을 출력하고, 회귀 분석은 두 변수의 관계를 설명하는 회귀식 정보들을 제공한다. 회귀식, 예측치, 잔차등의 전문 내용까지 출력되며, 두 개 이상의 독립변수를 가진 다중 회귀 분석 결과도 제공하므로 전문적인 회귀 분석 결과들을 모두 얻을 수 있다 (그림 3). [2,3]



<그림 3> 회귀분석 메뉴 대화상자

두개 이상 그룹 평균을 비교하는 분산분석에는 변수가 한 개인 경우에 사용할 수 있는 일원배치법 및 두개의 변수인 자료에 사용할 수 있는 이원배치법 메뉴가 엑셀에 존재한다. 전문 통계 패키지를 사용하는 경우 정해진 모형에 의하여 프로그램 코드를 작성해야 하는데, 엑셀에서는 자료 구성에 따라 대화상자에 필요한 값들만을 입력하면 되므로 매우 편리하게 원하는 결과들을 얻을 수 있다. (그림 4) [4]



<그림 4> 이원배치법 메뉴 대화상자

2-2. 자료수집

지하철 1,2,3,4 호선을 운영하는 서울메트로 홈페이지에서 역별 시간대별 지하철 승하차 인원 자료 파일, 역별 월평균 수입금 자료 파일 및 호선별 전력소모량 자료파일들에서 분석에 필요한 정보들을 추출하였다.

엑셀 파일들 이므로 정렬, 합병등의 기본데이터 메뉴에 의하여 가설검정에 필요한 변수들만을 쉽게 선택할 수 있었다. [5]

2-3. 연구 가설 및 모형

본 연구에서 분석하려는 내용은 승차 인원, 하차 인원이 운송 수입에 미치는 관계 검정, 매표소 개수에 따른 수입금과 승차 인원의 비교검정 및 호선과 월(月)이 전력소모량에 미치는 요인 검정이다.

변수들의 관련성 검정은 회귀 분석 모형, 비교검정은 t 검정 및 요인 분석을 위하여는 분산분석 모형을 적용하였다.

2-4. 엑셀 데이터분석 메뉴

양적 변수 기술을 위하여는 기술통계량 메뉴를, 질적 변수 기술을 위하여는 히스토그램 메뉴를 적용하였다. 변수들 관계 검정은 회귀분

석 메뉴를, 차이성 유무는 t검정 메뉴를, 요인 검정은 분산배치 메뉴를 적용하였다.

3. 결과

3-1. 기술통계

운송수입, 승하차 인원, 전력 소모량의 양적변수 및 매표소 개수의 질적변수가 표 1 과 표 2 에 기술되어있다.

<표 1> 양적 변수들의 기술통계량

승차		하차		수입	
평균	25085.09	평균	25214.84	평균	18577.34
표준 오차	1704.45	표준 오차	1705.176	표준 오차	1316.966
중앙값	20620	중앙값	21106.5	중앙값	14925
표준 편차	18357.49	표준 편차	18365.31	표준 편차	14184.16
분산	3.37E+08	분산	3.37E+08	분산	2.01E+08
첨도	1.670717	첨도	2.045071	첨도	2.090377
왜도	1.239635	왜도	1.273493	왜도	1.357919
범위	96593	범위	102044	범위	76254
최소값	617	최소값	841	최소값	379
최대값	97210	최대값	102885	최대값	76633
합	2909871	합	2924921	합	2154971
관측수	116	관측수	116	관측수	116

<표 2> 호선별 매표수 개수별 분포

계급	빈도수	누적 %	계급	빈도수	누적 %
1호선1개	2	1.72%	2호선2개	26	22.41%
1호선2개	8	8.62%	2호선1개	24	43.10%
2호선1개	24	29.31%	3호선1개	21	61.21%
2호선2개	26	51.72%	4호선2개	15	74.14%
3호선1개	21	69.83%	4호선1개	11	83.62%
3호선2개	9	77.59%	3호선2개	9	91.38%
4호선1개	11	87.07%	1호선2개	8	98.28%
4호선2개	15	100.00%	1호선1개	2	100.00%

3-2. 회귀 분석 결과

승차 인원과 하차인원의 독립변수, 운송수입의 종속변수를 가진 회귀 분석 결과는 표 3 에 있다. 추정된 회귀 모형은

Y=-571.7 + 0.62*X1 + 0.14 * X2

(Y: 운송 수입, X1: 승차인원, X2: 하차인원) 이었다.

승차인원과 운송수입과는 관계가 있었으나, 하차인원과 운송수입과는 관계가 없었다.

<표 3> 회귀분석 결과

요약 출력

회귀분석 통계량	
다중 상관계수	0.985808
결정계수	0.971818
조정된 결정계수	0.971319
표준 오차	2402.157
관측수	116

분산 분석

	자유도	제곱합	제곱 평균	F 비
회귀	2	2.25E+10	1.12E+10	1948.306
잔차	113	6.52E+08	5770359	
계	115	2.31E+10		

	계수	표준 오차	t 통계량	P-값
Y 절편	-571.725	380.0283	-1.50443	0.135261
승차인원	0.623981	0.091787	6.798168	5.26E-10
하차인원	0.138666	0.091748	1.511381	0.133483

3-3. 비교 분석 결과

매표소 1개 와 2개 의 운송수입 비교 결과가 표 4 에 있다. 매표소가 두개인 그룹 운송수입이 매표소가 한개인 그룹 운송 수입보다 높음이 통계적으로 증명되었다. ($P<0.01$)

<표 4> 이원분산분석 결과

t-검정: 이분산 가정 두 집단

	변수 1	변수 2
평균	39910.2069	60689.55172
분산	1075377015	1413653682
관측수	58	58
자유도	112	
t 통계량	-3.17198205	
P(T<=t) 단측	0.000976979	
P(T<=t) 양측	0.001953958	

3-4. 요인 검정 결과

계절과 호선에 따라 전력소비량에는 차이가 있음이 증명되었다 (표 4).

<표 4> 이원분산분석 결과

분산 분석

변동의 요인	제곱합	자유도	제곱 평균	F 비	P-값
인자 A(행)	0.890758	3	0.296919	7.705105	0.000518
인자 B(열)	3.054075	3	1.018025	26.41791	8.66E-09
교호작용	0.511025	9	0.056781	1.473464	0.200013
잔차	1.233133	32	0.038535		
계	5.688992	47			

그러나 계절과 호선의 동시 효과는 통계적으로 증명되지 않았다.

참 고 문 헌

[1] http://kin.naver.com/open100/db_detail.php?d1id=11&dir_id=1104&eid=gigyfo85vUZToBoJoF8XpNAdHdzJnfU8&qb=seK8+sXrsOg=&pid=fTa6Awoi5U4sstdO7CNsss--117947&sid=SWFKi7M-YUkAABQdGjY. 기술통계의 이론과 활용, 2008.10.

[2] http://kin.naver.com/detail/detail.php?d1id=11&dir_id=110203&eid=d+Swf7z+il/Ti1UXgdUbwvvHJvMLZZWG&qb=yLixzbrQvK4=&pid=fTaL1soi5T8ssab9n9hsss--264541&sid=SWFKi7M-YUkAABQdGjY. t 검정 결과와 회귀 분석 결과 차이점, 2008.12.

[3] http://bomool.net/bbs/board.php?bo_table=bo0107&wr_id=6. 엑셀에서 t검정, 2008.6.

[4] http://blog.naver.com/entopic?Redirect=Log&logNo=60036073662. 엑셀 2요인 분석, 2007.4.

[5] http://www.seoulmetro.co.kr. 자료실, 2008.

김숙영

미국 오하이오 주립대학 컴퓨터 과학과 졸업
미국 오하이오 주립 대학원 석사 (응용통계학)
현재 안산공과대학 컴퓨터정보과 재직

관심분야: 전산통계, 전산회계, 전산수학